

A Novel Universal Statistic for Computing Upper Limits in Ill-behaved Background

V. Dergachev¹

¹*LIGO Laboratory, California Institute of Technology, MS 100-36, Pasadena, CA 91125, USA*

(Dated: August 10, 2012)

Analysis of experimental data must sometimes deal with abrupt changes in the distribution of measured values. Setting upper limits on signals usually involves a veto procedure that excludes data not described by an assumed statistical model. We show how to implement statistical estimates of physical quantities (such as upper limits) that are valid without assuming a particular family of statistical distributions, while still providing close to optimal values when the data is from an expected distribution (such as Gaussian or exponential). This new technique can compute statistically sound results in the presence of severe non-Gaussian noise, relaxes assumptions on distribution stationarity and is especially useful in automated analysis of large datasets, where computational speed is important.

PACS numbers: 07.05.Kf, 02.50.Tt, 04.80.Nn, 06.20.Dk

INTRODUCTION

Data collected in experiments is sometimes contaminated by noise with an ill-behaved and often unknown distribution, presenting problems for the traditional method of using distribution quantiles to establish upper limits or confidence intervals. This problem happens especially often in experiments that collect large volumes of data.

A common solution is to exclude contaminated data from the analysis. For example, figure 1 shows a small portion of data obtained in the LIGO search for continuous gravitational waves in fifth science run. The blue points mark regions where non-Gaussian behaviour was detected and upper limit values are not expected to be valid.

In fact, if one looks carefully at the data for each point one can find a cause of non-Gaussian behaviour and a workaround to establish an upper limit - but the causes are different for different points, making analysis very laborious.

What is desired is an automated way to establish an upper limit that would be correct (if a bit conservative) for an arbitrary distribution, while still being close to optimum in the case of Gaussian noise (or other distribution class) that commonly occurs in the data.

We present a new algorithm that establishes upper limits without assuming a specific underlying background distribution, and that can be optimized for an arbitrary class of distributions (such as Gaussian, exponential, etc) that are expected to commonly occur in the data.

This advance allows one to obtain valid upper limits on signal strengths in the presence of ill-behaved and poorly understood background.

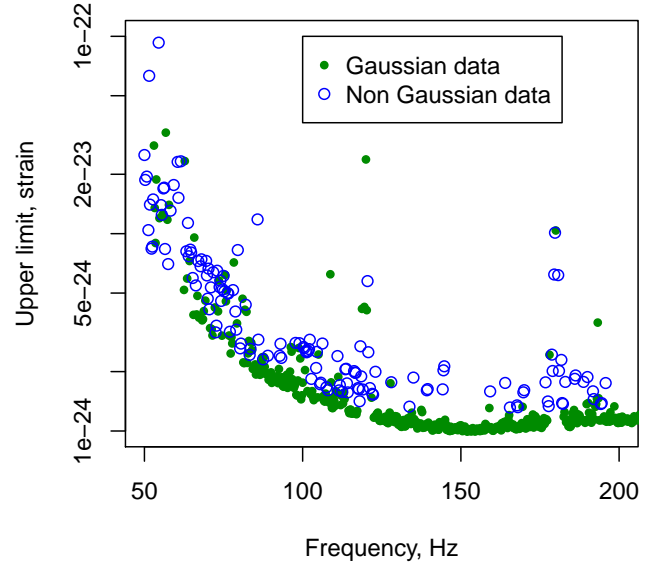


FIG. 1. Upper limit data from LIGO S5 search for continuous gravitational waves in the 50-200 Hz frequency range[1]. A large number of non-Gaussian bands significantly reduces the usefulness of the results. (color online)

UNIVERSAL INEQUALITIES AND STATISTIC

Let us consider a sample problem. Suppose we have obtained many samples of data which consists of background noise plus a possible signal. The data is collected in batches of N samples x_i for which the background noise ξ_i is independent and identically distributed. Also, we expect that at most one sample j in each batch contains a signal:

$$x_i = \xi_i + s\delta_{ij} \quad (1)$$

We would like to place a limit on the strength of the signals that may (or may not) be present in our data set.

If we knew that the noise ξ_i were drawn from a particular distribution ρ (such as Gaussian) our task would be straightforward - we would find the maximum x_i in all our data, subtract the mean of the background μ and add a distribution specific correction C_ρ that accounts for the possibility that a particular sample with the signal was below the background mean:

$$\text{UL}_\rho = \max_i x_i - \mu + C_\rho \quad (2)$$

If the distribution of ξ_i is not known with certainty, then we can try to estimate the distribution from the data itself. This, however, is problematic when the amount of data is small.

We now notice that, regardless of the procedure to compute the correction, it ends up being the function of the input data:

$$\text{UL} = \max_i x_i - \mu + C(\{x_i\}) \quad (3)$$

We can now pose the following problem:

Suppose we are given a confidence level $1 - \epsilon$, a class of commonly encountered distributions \mathcal{D} and a tolerance α . We need to find a function $C(\{x_i\})$ of the input data such that:

1. For any s and any distribution of ξ_i we have

$$P(\text{UL} < s) < \epsilon \quad (4)$$

2. We require that for any distribution $\rho \in \mathcal{D}$ the upper limits are overestimated by at most α compared to what we could obtain with full knowledge of the distribution:

$$\frac{\text{UL}}{\text{UL}_\rho} \leq 1 + \alpha \quad (5)$$

We call such statistics *universal* as they are applicable regardless of the distribution of noise we have.

DERIVATION OF UPPER LIMIT STATISTIC

In probability theory distribution-independent bounds are commonly obtained by use of Chebyshev-Bienaymé's or Markov's inequalities, however they are rarely used in practice, since in common applications they provide bounds that are far too loose.

For example, Encyclopaedia Britannica writes "Unfortunately, with virtually no restriction on the shape of an underlying distribution, the inequality is so weak as to be virtually useless to anyone looking for a precise statement on the probability of a large deviation. To achieve this goal, people usually try to justify a specific error distribution, such as the normal distribution..." [2].

This is because even though Chebyshev-Bienaymé's or Markov's inequalities are sharp - turning into equalities for an appropriate probability distribution - these distributions are rarely encountered in practice.

There exists a stronger Vysochanskij-Petunin inequality [3] but it relies on distributions being unimodal - an assumption that is hard to establish in empirical data. A review of other Chebyshev type inequalities can be found in [4, 5].

We engineer an upper limit statistic by starting with Markov's inequality

$$P(|X| \geq a) \leq \frac{\mathbb{E}|X|}{a} \quad (6)$$

and modifying it to read

$$P(|f(X)| \geq a) \leq \frac{\mathbb{E}|f(X)|}{a} \quad (7)$$

Then a further modification yields:

$$P\left(\left|f\left(\frac{X-\mu}{\sigma}\right)\right| \geq a\right) \leq \frac{\mathbb{E}\left|f\left(\frac{X-\mu}{\sigma}\right)\right|}{a} \quad (8)$$

for, in general, arbitrary μ and $\sigma > 0$ - though in practice these are chosen to be estimates of the mean and standard deviation. After setting

$$a = \frac{\mathbb{E}\left|f\left(\frac{X-\mu}{\sigma}\right)\right|}{\epsilon} \quad (9)$$

we obtain

$$P\left(\left|f\left(\frac{X-\mu}{\sigma}\right)\right| \geq \frac{\mathbb{E}\left|f\left(\frac{X-\mu}{\sigma}\right)\right|}{\epsilon}\right) \leq \epsilon \quad (10)$$

Because the original Markov's inequality is correct for a random variable X with an arbitrary distribution, inequality (10) is valid for any choice of $f(x)$, μ and σ - even when μ and σ are estimated from the data X .

We can now optimize $f(x)$ to provide more precise upper limits or confidence intervals for our desired distribution. As a quick example, the inequality 10 becomes sharp for a Gaussian random variable X when we choose $\mu = \mathbb{E}X$, $\sigma = \sqrt{\text{Var } X}$ and use a step function

$$f_s(x) = \begin{cases} 1 & \text{when } x \geq \hat{x}_\epsilon \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

where \hat{x}_ϵ satisfies

$$\mathcal{F}(\hat{x}_\epsilon) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\hat{x}_\epsilon} e^{-x^2/2} dx = 1 - \epsilon \quad (12)$$

The choice $f(x) = f_s(x)$ is difficult to apply to establish a confidence interval because the function $f_s(x)$ is not invertible: it can happen that the average of $\frac{1}{\epsilon} \left|f\left(\frac{X-\mu}{\sigma}\right)\right|$

for practical data is greater than 1 which does not yield a constraint on X . One approach could be to pick initial x_ϵ as defined by equation 12 and then iterate to establish a bound for X . This is cumbersome for both analytical and numerical computation.

A better way is to use an $f(x)$ that is invertible above x_ϵ . An especially simple and computationally efficient example, shown in figure 2, is given by

$$f_c(x) = \begin{cases} 1 + \frac{1}{2}(x - x_\epsilon) & \text{when } x \geq x_\epsilon \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

with the corresponding inverse function given by

$$f_c^{inv}(x) = \begin{cases} x_\epsilon + 2(x - 1) & \text{when } x \geq 1 \\ x_\epsilon & \text{otherwise} \end{cases} \quad (14)$$

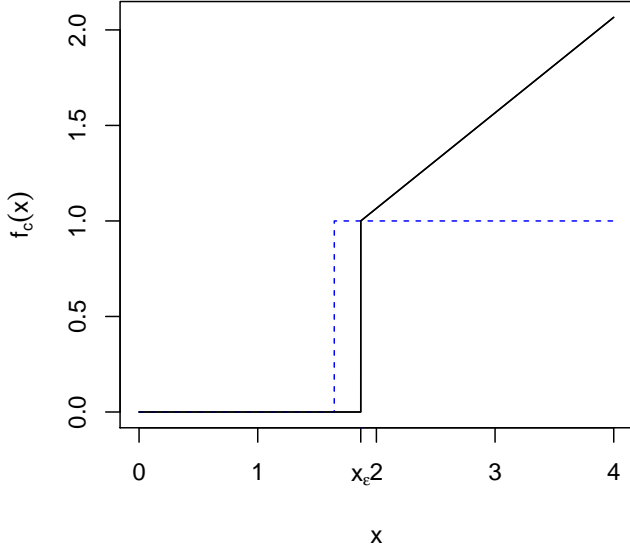


FIG. 2. Function f_c used for computation of 95% confidence level upper limits (simulation results are shown in figure 4). The point x_ϵ has been shifted to the right to compensate for errors in mean estimates for 501 points of input data. The dashed line shows the step function f_s that makes inequality 10 exact in the ideal case. (color online)

Our correction C is then

$$C = \sigma f_c^{inv} \left(\mathbb{E} \left(\left| f_c \left(\frac{X - \mu}{\sigma} \right) \right| \right) \right) \quad (15)$$

with expectation replaced by average for empirical data.

PERFORMANCE OF UNIVERSAL UPPER LIMIT STATISTIC

A step by step algorithm for computing the upper limit is shown in figure 3. It incorporates two adjustments

that we find important in practical implementation.

First, the point x_ϵ has been increased by $5/\sqrt{N}$ to compensate for possible errors in estimation of mean μ which could lead to effectively underestimating x_ϵ .

Secondly, the standard deviation σ is estimated using data from the lower tail of the distribution only as appropriate for establishing an upper limit.

The main steps 2-6 of the algorithm 3 employ only piecewise linear functions allowing for very efficient implementation on virtually any computational platform.

1. Prepare by computing value of

$$x_\epsilon = \mathcal{F}^{-1}(1 - \epsilon) + 5/\sqrt{N}$$

2. Compute $M = \max_{i=1..N} x_i$
3. Compute $\mu = \frac{1}{N} \sum_{i=1}^N x_i$.
4. Compute $\sigma = -\frac{\sqrt{2\pi}}{N} \sum_{i=1}^N \min(x_i, 0)$. We prefer this formula as it is simpler to compute and is insensitive to outliers in the upper tail of the distribution.
5. Compute $\delta = \frac{1}{N} \sum_{i=1}^N f_c \left(\frac{x_i - \mu}{\sigma} \right)$.
6. Establish upper limit $UL = M - \mu + \sigma f_c^{inv}(\delta)$

FIG. 3. Algorithm for computing the upper limit from a single batch of data of N points.

To gauge the performance of our new statistic, we performed a simulation that closely reflects real-world situations we encountered in data analysis [1]. We assume that our data consists of independent samples of noise plus a possible deterministic signal in one or more bins. The data is analyzed in batches of 501 data samples for each of which we establish an upper limit on signal strength. The final reported value is the worst case (i.e. maximum) upper limit among 100 batches.

The comparison of this worst case upper limit to the upper limit established analytically from the known distribution of underlying noise is shown in figure 4. The samples consisted of identically distributed pure noise ($s = 0$).

We have averaged the ratios of established upper limits to theoretical ideal values. As the value of x_ϵ was obtained assuming Gaussian data we see that our statistic achieves less than 5% overestimate both for 90% and 95% confidence level upper limits.

A number of other distributions have been tried. As seen on the plot, the performance is remarkably flat for χ^2 distributions with different degrees of freedom and the overestimate is moderate for uniform distribution. The heavy-tailed Student's t-distributions, as well as lognormal distribution, show good performance as well. Even in the extreme case of Bernoulli distribution, with equal probability to obtain 0 and 1, our overestimate is less than 50% for 95% confidence level. Finally, the custom distribution `test1` composed of three populations of nor-

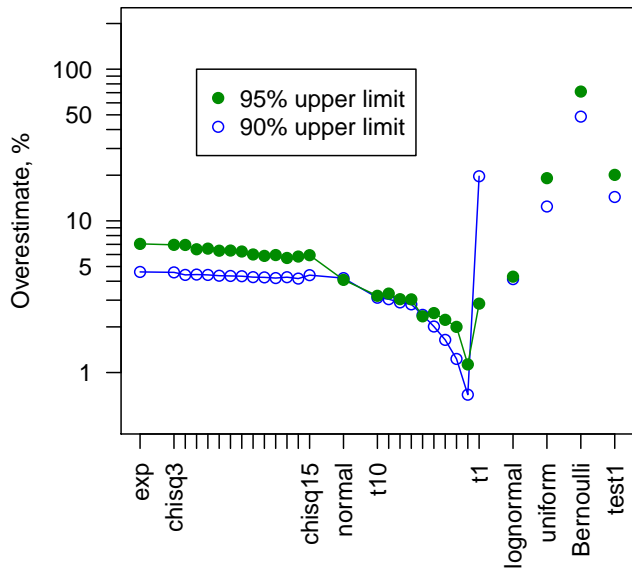


FIG. 4. Average overestimate of upper limit by the universal statistic as compared to the value predicted by analytical formula for the corresponding distributions. The overestimate is less than 5% for Gaussian data, and we expect any practical measurement to perform worse than the ideal case. The upper limits were computed using f_c (see figure 2) with 501 points of data for different noise distributions: exponential, χ^2 with different degrees of freedom, Gaussian, Student’s t-distribution with different degrees of freedom, lognormal, uniform, Bernoulli with equal probability to obtain each outcome and a custom distribution **test1** the histogram for which is shown on figure 5. The points on the graph were obtained by averaging 100 independent measurements, each of which consisted of finding the maximum among 100 upper limits to simulate maximization across a set of templates. (color online)

mal and exponentially distributed numbers (figure 5) has overestimate of only 20% for 95% confidence level.

CONCLUSIONS

We have described a new *universal* statistic that produces reliable and useful upper limits regardless of the underlying distribution of noise, while still producing close to optimum values for a specific family of distributions. The algorithm for computing its values is very practical, and is easily implemented for large scale computation.

This opens the road for publication of reliable results from large data sets with only partial understanding of distributional properties of data they contain.

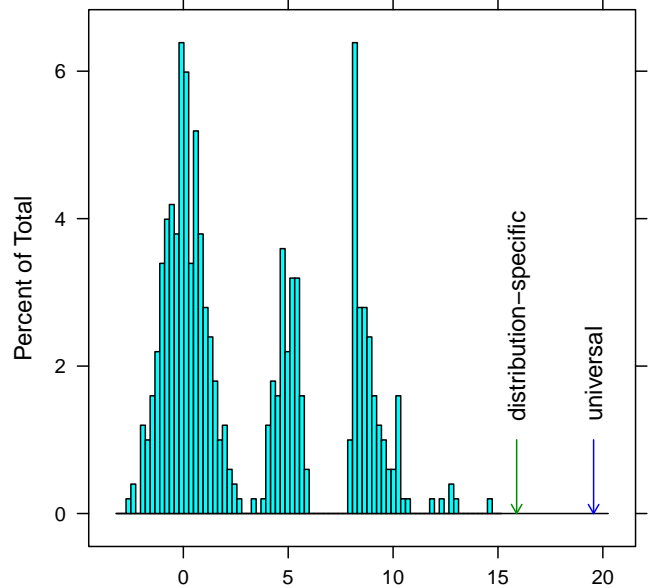


FIG. 5. Distribution **test1** used in figure 4. It is composed of three populations, two normal and one exponential. We also show distribution-specific and universal 95% confidence level upper limit for this batch of 501 numbers. (color online)

ACKNOWLEDGMENTS

This work has been done while being a member of LIGO laboratory, supported by funding from United States National Science Foundation. LIGO was constructed by the California Institute of Technology and Massachusetts Institute of Technology with funding from the National Science Foundation and operates under co-operative agreement PHY-0757058.

The author has greatly benefited from suggestions and comments of his colleagues, in particular Keith Riles, Evan Goetz and Roy Williams.

This document has LIGO Laboratory document number LIGO-P1200065-v4.

-
- [1] B. Abbott *et al.* (The LIGO and Virgo Scientific Collaboration), *Phys. Rev.* **D 85**, 022001 (2012)
 - [2] Chebyshev’s inequality. (2012). In *Encyclopaedia Britannica*. Retrieved from <http://www.britannica.com/EBchecked/topic/108218/Chebyshevs-inequality>.
 - [3] D. F. Vysochanskij, Y. I. Petunin, *Theory of Probability and Mathematical Statistics* **21**: 2536 (1980).
 - [4] F. Pukelsheim *The American Statistician*, **48**, No. 2 (May, 1994), pp. 88-91
 - [5] I. R. Savage, *Journal of Research of the National Bureau of Standards - B.Mathematics and Mathematical Physics*,

65 B, 211-222.